SHIV SHAKTI
International Journal in Multidisciplinary and Academic Research (SSIJMAR)
Vol. 1, No. 3, September-October (ISSN 2278 - 5973)

ANALYTICAL ANALYSIS of CLUSTERING ALGORITHMS TO BOOST THE COMPETENCE OF THE NETWORK INTRUSION DETECTION SYSTEM

Sonam Lowry*

Abstract

Clustering is an significant task in mining evolving data streams. Most current intrusion detection systems are signature based ones or machine learning based methods. The research and investigation of the intrusion detection techniques and data mining, this paper presents a intrusion detection model which is based on cluster analysis K-means algorithm.

Index Terms— intrusion detection; data mining; Clustering Analysis

*Sonam Lowry is with the Computer Engineering Department, Poornima college of Engineering. Sitapura, Jaipur 302022 INDIA (e-mail: er_sonamlowry@yahoo.co.in).

INTRODUCTION

Intrusion detection systems (IDS) are safety tools designed
to detect and classify attacks against computer networks
and hosts. They can drive in two ways: by searching for specific patterns in data (misuse based IDS) or by recognizing certain deviations from expected behavior (anomaly based IDS). In anomaly based IDS, clustering algorithms are often used for recognition of"abnormal" behavior. They can be applied either directly on incoming data [1,2] or as a supporting technique in a stage posterior to data classification performed by means of other techniques .Anomaly based IDS classify input data into a number of categories, or classes. This number can be arbitrary, but as the essential goal of these systems is to distinguish between "normal" and "abnormal" behavior , it is very common to partition the incoming resource access requests into two classes that correspond to these two types of behavior. As the size of computer networks increase rapidly, the network security issues become increasingly prominent. As a proactive defense technology, Intrusion detection is expected to provide achieve all-round test for network attacks. The essence of the network intrusion detection is the analysis the data which is from the network to the data, so the use of data mining technology for the application of intrusion detection research has been extensive attention at home and abroad.

Intrusion detection technology is divided into two types: misuse detection and anomaly detection. Misuse detection is to collect information, then match it with the known intrusion pattern, the main methods is expert systems, state transition analysis and so on. The professional systems are based on rules. The advantage of misuse Detection system is the burden on small, high rate of detection accuracy, of the intrusion detection products are based on the misuse detection, but the weakness of this method is not found in unknown attacks. Anomaly detection assume that the incursion activities are distinct from the main customary activities, establishing "activity overview" when the structure is under normal operation, then as it a basis to find the behaviors those who deviate from the normal activity pattern. Early abnormal detection is mainly statistical analysis method, for example: variance analysis, Bayesian classification, Markov process, etc, the disadvantage of this approach is that loss detecting rate and false detection rate of intrusion detection is incredibly high. The data sources of intrusion detection are mainly network data packets and host audit logs, these data are very large, in order to discover potentially useful knowledge from this wealth of information, so must rely on data mining of this powerful tool.

Intrusion Detection System (IDS) has turn out to be an important study area in Computer based security [3]. It is a renowned skill for illuminating and is used as a countermeasure to protect data reliability and system ease of access during an intrusion [4]. When a client tries to access into an information structure or carries out an action illegally, the action is referred as an intrusion that can be separated into two sets, *exterior and interior*. The exterior refers to those clients who do not have authorized access to the system and who tried to access illegally by using different saturation methods. Whereas interior refers to those which have valid access permission but desire to carry out illegal activities. Methods for the intrusion may include software bugs exploitation and misconfigurations of the system, password cracking, sniffing unsecured traffic, or utilizing the specific protocols design flaw.

I. THE DESCRIPTION OF DATA MINING

A. Data Mining Technology

The definition of Data Mining has two kinds: broad and narrow sense. From the extensive point of view, Data mining is from bulky data sets (which may be incomplete, noise, uncertainties, and various forms of storage), mining implicit in the sets, that people do not know in advance, and useful knowledge for decision-making process. From the narrow point of

view, data mining can be defined a process, that is refine knowledge from a particular form of data-intensive. There are numerous data mining methods, according to the diverse mining tasks, data mining can be alienated into four types: association analysis, sequence analysis, classification analysis and cluster analysis.

B. Distance Definition

Clustering is that corporeal or abstract objects together into a groups formed by similar pairs of manifold classes of process. The difference with classification is that it wants to divide the class, but the class is unknown, among the same class of objects, they have greater similarity, however, among the different types of objects they are moderately different. Similarity is used to describe the degree of similarity between two objects.

III. Intrusion Detection Model

For a successful intrusion detection system, it is not only to make system administrators know the network systems (including procedures, documentation and hardware equipment, etc.) of any changes, but also provide guidance to the formulation of the network security policy. More outstandingly, it should be management, pattern simple, so that non-professionals are exceptionally easy access to network security. Moreover, the scale of intrusion detection should be changed on the basis of network threats and security needs of system structure. Intrusion detection system will be a timely response after finding the invasion, including the disconnection of network connectivity, recording events and alarms. The key of anomaly detection technology is how to construct a normal behavior network model, and the cluster analysis in data mining provides an effective way to solve this problem. The advantage of cluster analysis algorithm applied to the abnormal detection is that it can automatically construct the network model of normal behavior, and it is without manual operation. In addition, abnormal detection based engine based on cluster analysis can also play a role as filters. It can be a normal packet filter, so that you can reduce unnecessary rules that match the job. Under regular circumstances, most of the network data packet is well-balanced; therefore, abnormal detection model can increase the speed of processing data packets. The network data packets which do not meet the normal behavior model are regarded as abnormal data packets; they are further tested by misuse detection. Those nonstandard data packets which misuse detection engine did not find is likely to be generated by the new intrusion packets. Associating packets of these anomalies with the invasion analysis, we can get new patterns of behavior, then change the intrusion behavioral patterns for the
intrusion detection rules and added to the rule base, so misuse detection engine can detect the new intrusion. According to the above design idea, this article makes data mining technology applied to network intrusion detection systems, on the basis of the original model. It increases the inventive cluster analysis module, anomaly detection, correlation analyzer, and constructs Network Intrusion Detection System Model which is based on data mining.

A. Cluster analysis

Cluster scrutiny module constructs a mold of normal behavior. Cluster analysis module is called after the anomaly detection component and misuse detection module. It guarantees that cluster analysis will not repeat producing the normal behavior model, it can also ensure that Cluster analysis can not manufacture known intrusion model. Because after anomaly detection module and misuse detection modules filtering, the data packet enters into the cluster analysis module, but packet is a normal behavior unknown or unknown intrusion generated. Cluster analysis module will add the new network normal behavior models to the abnormal detection module, the abnormal data packets are failure to form a network model of normal behavior, so they will be recorded in an exception log. As the network intrusion detection system demands high real-time, therefore, cluster analysis module must adopt an efficient clustering analysis algorithm,

otherwise, when the network traffic is larger, system is prone to the observable fact of packet loss, which would have omitted. K-Means algorithm is simple and its computational complexity is smaller, therefore, K-Means algorithm for the clustering analysis algorithm.

B. Abnormal detection

Abnormal detection [4] technology assumes that all of the intrusions must be abnormal activity, in this case, if establish a normal activity for the system characteristics of the file (Profiles), in theory, stat the distinctiveness of the file number of the state of all systems, and they are different from state it has established, then, identify intrusion attempts through them. In fact a collection of invasion activity is not necessarily a collection of unusual activity;
the condition as follows:

①Abnormal activity is not a invasion activity, but it
has been identified for the invasion, which we call false
positives, and it will cause false alarms.
②incursion activity was not strange activity, it is that
invasion activity is identified as the normal activities,
which call omitted, this will result missed sanction, it
will be more serious than the first case.
The key of the abnormal detection question is how to
select the proper hold value, which can make the
above two kinds of situation not undue expansion, another
thing is how to select the measurable characteristics you
want to monitor.

C. Association analysis

Association analysis is a simple and practical correlation analysis rule, when some of the attributes of a transaction occurs at the same time, it describes the laws and models, and it is a data segmentation technology by a series of "if ~ then" logical rules. Association rules in general relate to the operation database, and each transaction consists of a group of records. This operation database usually includes very large data; therefore, according to a certain degree of support count of the record, the current association rule discovery technology is trying to reduce the look at space. The degree of support is a metric which is a transaction number appearing in the log. Association analysis can be divided into two steps [5]:
①Generating recurrent item sets: frequent item sets can meet the minimum support.
②Generating well-built association rules: the strong association rules can meet the smallest degree of confidence.

IV. IMPROVED ALGORITHM

A. Clustering Analysis Algorithm

Clustering analysis algorithm [6] is that: first it compares variables and then attributes the data with similar characteristics to a class. Thus, by clustering, the data set translate into a class set, in the class set, the same type of data has a similar variable value, among different types of data variable value does not have similarities. How to differentiate different classes is a part of the data mining process, these classes are not pre-defined, these classes by using fully automatic manner through the clustering. What sort of cluster analysis algorithm is used depends on the form of data, clustering the purpose and application. The main clustering methods are: hierarchical methods, partition method, density-based methods, grid based methods and model-based methods. Attribute-Based Clustering intrusion detection algorithm is based on two assumptions: first, the normal behavior data is far greater than the intrusion data; second, the invasion behavior varies widely with normal behavior. Partition method is used the most widely; partition method can be divided into K-mean (K-means) algorithm and the K-center (Kmedoids) algorithm. Secretarial the specific reality, we adopt

the k-means [7] algorithm. the specific algorithm is as follows:

Input: The number of clusters k and a Sample set containing n objects.

Output: A set of k clusters that minimizes the square error criterion.

Method:

（1）randomly choose k objects as initial cluster centers;

（2）sequence (3), (4), until each cluster will not change;

（3）assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

（4）inform the cluster means, i.e., is calculated the mean value of the objects for each cluster;

（5）until the cycle no changed.

B. Enhanced Algorithm

In the intrusion detection, we apply k-means algorithm, we found that different initial cluster centers will result in different effects for the final test. But in the original kmeans algorithm, the choice of the initial cluster centers is arbitrary, so before trying to apply the algorithm we hope that we can get better initial cluster center. Improved algorithm described as follows:

Suppose the U sample sets have n samples, cluster it into k classes, and the initial value of m is 1.

1) count up the distance $d(X, Y)$ between the two samples, find the nearest two points in the sample set U , forming collection $Am(m \in [1, k])$. Points from the sample set U ;

2) Find the nearest sample point from the Am in U , access it to Am, and then remove it from the U ;

3) Repeat Step 2), until the collected works of sample points *Am* account a certain proportion of □ in all sample points;

4) If $m \neq k$, then m□ according to 1),2),3)step, operating the U removing from the original sample set U , and get new *Am* ;

5) After a number of iterations to find, get k-set, seek average for each set of samples, then define The average value as the set of initial clustering centers;

6) Now the original k-mean algorithm to cluster.

V. CONCLUSION

The main intend of intrusion detection system in this paper is to improve the data analysis module. It will be a effective intrusion detection for unknown intrusion and composite intrusion action by using k-means algorithm, it can reduce the error rate; and it will achieve better test results. In this paper, it was only a preliminary attempt by combining data mining technologies and clustering algorithm to the network intrusion detection system for the study.

REFERENCES

[1] Frank J., ”Artificial Intelligence and Intrusion Detection:
Current and Future Directions”, *Proceedings of the 17th*
National Computer Security Conference, Baltimore, USA,

[2] Guan Y., Ghorbani A. and Belacel N., ”Y-Means: a
Clustering Method for Intrusion Detection”, *Proceedings*
of Canadian Conference on Electrical and Computer
Engineering, Montreal, Canada, 2003.

[3] McHugh, John, 2001. “Intrusion and Intrusion Detection.”
Technical Report. CERT Coordination Center, Software
Engineering Institute, Carnegie Mellon University.

[4] D. Wagner and D. Dean, “Intrusion detection via static
analysis,” in Proc. IEEE Symposium on Research in Security and Privacy, Oakland, CA, 2001.

[5]Zhu’an qing, Zhang chang cheng. “Research on network intrusion
detection technology based on data mining” [D]. Computer
Engineering and Design, 2008.

[6] Lixiaorui. “Research on Database Intrusion Detection”[DI].
Beijing:Beijing Polytechnic University, 2008.

[7] Jiaweihan. “ Data mining” [M]. Higher Education Press, 2003.